U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
NATIONAL METEOROLOGICAL CENTER

OFFICE NOTE 382

"APPLICATION OF THE GANDIN-MURPHY EQUITABLE SKILL SCORE TO NUMERICAL FORECASTS OF QUANTITATIVE PRECIPITATION"

JOSEPH P. GERRITY, JR.
DEVELOPMENT DIVISION

JUNE 1991

THIS IS AN UNREVIEWED MANUSCRIPT, PRIMARILY INTENDED FOR INFORMAL EXCHANGE OF INFORMATION AMONG NMC STAFF MEMBERS

382

Q1 is the probability that the event is forecast to occur; and Q2 is the probability that the event is forecast not to occur.

Also use the notations:

Q1*P1 = Prob(event forecast and event observed)
Q1*P2 = Prob(event forecast and event not observed)
Q2*P1 = Prob(event not forecast and event observed)
Q2*P2 = Prob(event not forecast and event not observed)

In terms of these probabilities the expected form of the contingency table is:

		Forecast		
		Yes	No	Sum
	Yes	Q1*P1	Q2*P1	P1
Observed	No	Q1*P2	Q2*P2	P2
	Sum	Q1	Q2	1.

A "random" prediction is one in which no correlation exists between forecast and observation; the forecast and observation are statistically independent. In such a forecast process, one may write:

P1*Q1=(P1)(Q1), P2*Q2=(P2)(Q2), P1*Q1=(P1)(Q1) and P2*Q2=(P2)(Q2)

The entries in the contingency table can be displayed as an array C,

$$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

where

$$c_{11} = P1*Q1 \qquad c_{12} = P1*Q2$$

$$c_{21} = P2*Q1 \qquad c_{22} = P2*Q2$$

and then the expected value of the score S may be expressed as

$$E\{S\} = Trace\left(W^T C\right)$$

The <u>equitable</u> <u>score</u> was designed so that its "expected value" E{S} vanishes for the three "zero skill" forecasts:

#1. Q1 = 1 ; always forecast Yes
#2. Q2 = 1 ; always forecast No
#3. Random prediction

and has the value unity for a perfect forecast,

#4. H1 = O and H2 = NO

These criteria impose the conditions expressed by the equations,

$$w_{11}P1 + w_{21}(1.-P1) = 0 \qquad (2a)$$

$$w_{12}P1 + w_{22}(1.-P1) = 0 \qquad (2b)$$

$$Q1\{w_{11}P1 + w_{21}(1.-P1)\} + (1.-Q1)\{w_{12}P1 + w_{22}(1.-P1)\} = 0 \qquad (2c)$$

$$w_{11}P1 + w_{22}(1.-P1) = 1 \qquad (2d)$$

for all P1 and Q1.

As shown by Gandin and Murphy the conditions are satisfied when :

$$w_{11} = \frac{(1.-P1)}{P1} \qquad (3a)$$

$$w_{22} = \frac{P1}{1.-P1} \qquad (3b)$$

$$w_{12} = w_{21} = -1 \qquad (3c)$$

The vanishing of the expected value of S is not affected by a forecasting strategy intended to give preference to either the occurrence or non-occurence of the event. For any random forecasting system, irrespective of the "bias" (ratio of Q1 to P1), the random forecast gives the <u>equitable score</u>

$$S_R = (Q1\ \ P1)(P2/P1) + (Q2\ \ P2)(P1/P2) - (Q1\ \ P2) - (Q2\ \ P1) = 0 \qquad (4)$$

provided the weights (eq. 3) are computed using the probabilities (P1,P2) reflected in the sample underlying the contingency table.

On the other hand, the value of the <u>threat score</u> computed for the random forecast contingency table gives,

$$TS_R = \frac{(Q1)(P1)}{Q1 + P1 - (Q1)(P1)} \qquad (5)$$

Introducing the bias (B = Q1 / P1,) the <u>threat score</u> for a random forecast is,

$$TS_R = \frac{BP1}{(1.+B) - BP1}, \qquad (6)$$

which does not vanish. This implies (at least implicitly) that a random forecast possesses skill; further, this "skill" can be enhanced by using a biased estimate of the probability of occurrence of the event. As a consequence, both the Bias and the threat score need to be considered when assessing the performance of a forecaster or a forecasting system.

As a further clarification, recall that the day-to-day evaluation of the threat score can yield significant variations due to the fluctuation of O/N about its expected value P1. This can be seen in calculations of the correlation between the threat score and O/N, and has long been known to favor the threat score achieved by subjective forecasters who have the good fortune to work during periods of above normal occurrence of precipitation. This property of the threat score is compounded by the influence of a biased estimate of P1 on the score's non-zero, expected value for a random forecast.

If the equitable score is computed for a sample of forecasts, produced by a forecaster or forecasting system, it is important that the probabilities used in defining the weights be representative of the sample. Alternatively, the samples be sufficiently large that climatological probabilities are appropriate. Observing this precaution, it is to be anticipated that biased forecasting strategies will be ineffectual.

Gandin and Murphy outlined a comprehensive, but complex, method for determining the equitable score weights for multiple class contingency tables. A relatively simple method for generating the weights was found and applied to three and four class tables (sections 2 and 3, respectively.)

The results of computation of equitable score and other statistics for a set of May 1991 precipitation forecasts by the ETA model (Mesinger, Janjic and Black, 1990) are contained in section 4. A comparison between quantitative precipitation forecasts (June 1991) by the Nested Grid Model and the ETA model is made in section 5. Additional remarks on the equitable score appear in section 6; the paper ends with a summary and statement of some expectations for future work.

2 THREE CLASS TABLE

In this section, an <u>equitable</u> skill score is constructed for a contingency table with three classes of the event. The selected classes exhaust the set of possible outcomes and are ordinally related. Consider for concreteness the prediction of quantitative precipitation in one of three classes: Light, Moderate, Heavy. Suppose the contingency table has the entries:

		Forecast			
		light	moderate	heavy	sum
Observed	light	n11	n12	n13	p1
	moderate	n21	n22	n23	p2
	heavy	n31	n32	n33	p3
	sum	q1	q2	q3	N

To take advantage of Gandin and Murphy's two class equitable score weight matrix, partition the three-class contingency table into two, two-class tables:

Table 1.

		Forecast		
		light	moderate or heavy	sum
Observed	light	n11	(n12+n13)	p1
	moderate or heavy	(n21+n31)	(n22+n23+n32+n33)	(p2+p3)
	sum	q1	(q2+q3)	N

Table 2.

		Forecast		
		light or moderate	heavy	sum
Observed	light or moderate	(n11+n12+n21+n22)	(n13+n23)	(p1+p2)
	heavy	(n31 + n32)	n33	p3
	sum	(q1+q2)	q3	N

Using the definitions:

$$P1 = E\left\{\frac{p1}{N}\right\}, P2 = E\left\{\frac{p2}{N}\right\}, P3 = E\left\{\frac{p3}{N}\right\} \tag{7a}$$

$$Q1 = E\left\{\frac{q1}{N}\right\}, Q2 = E\left\{\frac{q2}{N}\right\}, Q3 = E\left\{\frac{q3}{N}\right\} \tag{7b}$$

the <u>equitable</u> weight matrix associated with 2-class Table 1 is

$$\begin{pmatrix} w_{11}^{(1)} & w_{12}^{(1)} \\ w_{21}^{(1)} & w_{22}^{(1)} \end{pmatrix}$$

with

$$w_{11}^{(1)} = (P2 + P3)/P1 \tag{8a}$$

$$w_{22}^{(1)} = P1/(P2 + P3) \tag{8b}$$

$$w_{12}^{(1)} = w_{21}^{(1)} = -1 \tag{8c}$$

and that associated with Table 2 is

$$\begin{pmatrix} w_{11}^{(2)} & w_{12}^{(2)} \\ w_{21}^{(2)} & w_{22}^{(2)} \end{pmatrix}$$

with

$$w_{11}^{(2)} = P3/(P1 + P2) \tag{9a}$$

$$w_{22}^{(2)} = (P1 + P2)/P3 \tag{9b}$$

$$w_{12}^{(2)} = w_{21}^{(2)} = -1 \tag{9c}$$

Evaluating the <u>equitable</u> <u>scores</u> for tables 1 and 2, and denoting them by S1 and S2, yields:

$$S1 = n11\,w_{11}^{(1)} + (n22 + n23 + n32 + n33)w_{22}^{(1)} + (n12 + n13)w_{12}^{(1)} + (n21 + n31)w_{21}^{(1)} \tag{10}$$

$$S2 = n33\,w_{33}^{(2)} + (n11 + n12 + n21 + n22)w_{11}^{(2)} + (n13 + n23)w_{12}^{(2)} + (n32 + n31)w_{21}^{(2)}$$

In terms of S1 and S2, <u>an</u> <u>equitable</u> score S, for the three class table, may be written

$$S = \alpha S1 + (1. - \alpha)S2, \tag{11}$$

where α is arbitary, but chosen to be .5 for specificity. The weight matrix implied by this definition of an <u>equitable</u> score may be worked out by noting the multipliers of each element of the three class table arising in the equation defining the scores S1 and S2. The result is the weight matrix

$$\begin{pmatrix} \hat{w}_{11} & \hat{w}_{12} & \hat{w}_{13} \\ \hat{w}_{21} & \hat{w}_{22} & \hat{w}_{23} \\ \hat{w}_{31} & \hat{w}_{32} & \hat{w}_{33} \end{pmatrix}$$

in which

$$\hat{w}_{11} = \frac{1}{2}\left(w_{11}^{(1)} + w_{11}^{(2)}\right) \qquad \hat{w}_{12} = \hat{w}_{21} = \frac{1}{2}\left(w_{12}^{(1)} + w_{11}^{(2)}\right) \tag{12a}$$

$$\hat{w}_{22} = \frac{1}{2}\left(w_{22}^{(1)} + w_{11}^{(2)}\right) \qquad \hat{w}_{13} = \hat{w}_{31} = \frac{1}{2}\left(w_{12}^{(1)} + w_{12}^{(2)}\right) \tag{12b}$$

$$\hat{w}_{33} = \frac{1}{2}\left(w_{22}^{(1)} + w_{22}^{(2)}\right) \qquad \hat{w}_{23} = \hat{w}_{32} = \frac{1}{2}\left(w_{22}^{(1)} + w_{21}^{(2)}\right) \tag{12c}$$

or in terms of the probabilities:

$$\hat{w}_{11} = \frac{(P2+P3)}{(2P1)} + \frac{(P3)}{2(P1+P2)} \qquad \hat{w}_{12} = \hat{w}_{21} = \frac{-1}{2} + \frac{(P3)}{2(P1+P2)} \tag{13a}$$

$$\hat{w}_{22} = \frac{(P1)}{2(P2+P3)} + \frac{(P3)}{2(P1+P2)} \qquad \hat{w}_{23} = \hat{w}_{32} = \frac{-1}{2} + \frac{(P1)}{2(P2+P3)} \tag{13b}$$

$$\hat{w}_{33} = \frac{(P1+P2)}{(2P3)} + \frac{(P1)}{2(P2+P3)} \qquad \hat{w}_{13} = \hat{w}_{31} = -1 \tag{13c}$$

To prove that the score S is equitable, it is necessary to show that it satisfies the following conditions:

1. S = 0. ; if Q1=1., or Q2=1., or Q3=1.

The contingency table for Q1=1 is

		Forecast			
		C1	C2	C3	Sum
Observed	C1	P1	0.	0.	P1
	C2	P2	0.	0.	P2
	C3	P3	0.	0.	P3
	Sum	1.	0.	0.	1.

The contingency table for Q2=1 is

		Forecast			
		C1	C2	C3	Sum
Observed	C1	0.	P1	0.	P1
	C2	0.	P2	0.	P2
	C3	0.	P3	0.	P3
	Sum	0.	1.	0.	1.

The contingency table for Q3=1 is

		Forecast			
		C1	C2	C3	Sum
Observed	C1	0.	0.	P1	P1
	C2	0.	0.	P2	P2
	C3	0.	0.	P3	P3
	Sum	0.	0.	1.	1.

2. S=0. ; if the forecast is randomly generated with arbitrary Q's; in this case the contingency table takes the expected form:

		Forecast			
		C1	C2	C3	Sum
	C1	Q1 P1	Q2 P1	Q3 P1	P1
Observed	C2	Q1 P2	Q2 P2	Q3 P2	P2
	C3	Q1 P3	Q2 P3	Q3 P3	P3
	Sum	Q1	Q2	Q3	1.

3. S = 1. ; the forecast is perfect, i.e. the expected three class contingency table has the form:

		Forecast			
		C1	C2	C3	Sum
	C1	P1	0.	0.	P1
Observed	C2	0.	P2	0.	P2
	C3	0.	0.	P3	P3
	Sum	P1	P2	P3	1.

For the case in which event 1 is always forecast, i.e. Q1=1, the score S is

$$S = P1\hat{w}_{11} + P2\hat{w}_{21} + P3\hat{w}_{31}$$

$$= .5\left[P1\left(w_{11}^{(1)} + w_{11}^{(2)}\right) + P2\left(w_{21}^{(1)} + w_{11}^{(2)}\right) + P3\left(w_{21}^{(1)} + w_{21}^{(2)}\right)\right]$$

$$= .5\left[\left\{P1\,w_{11}^{(1)} + (P2+P3)w_{21}^{(1)}\right\} + \left\{(P1+P2)w_{11}^{(2)} + P3\,w_{21}^{(2)}\right\}\right] \qquad (14)$$

Since S1 and S2 are equitable scores, it follows that the two factors enclosed in curly braces each vanish. Hence, S also vanishes when Q1=1., as required for S to be an <u>equitable</u> score. In the interest of space, we omit the similar proofs for Q2=1 and Q3=1.

The satisfaction of the condition 2. (random forecast) follows directly from the satisfaction of the three parts of condition 1. This is seen when the equitable score S, using the random forecast contingency table, is written:

$$\begin{aligned} S = Q1&\left\{P1\hat{w}_{11} + P2\hat{w}_{21} + P3\hat{w}_{31}\right\} \\ +Q2&\left\{P1\hat{w}_{12} + P2\hat{w}_{22} + P3\hat{w}_{32}\right\} \\ +Q3&\left\{P1\hat{w}_{13} + P2\hat{w}_{23} + P3\hat{w}_{33}\right\} \end{aligned} \qquad (15)$$

The coefficient of Q1 in equation (15) is the combination of terms which was shown to vanish in equation (14). The coefficients of Q2 and Q3 are similarly the combination of terms which must vanish for the two cases, Q2=1, Q3=1.

Finally, the proof that S has the value unity for a perfect forecast follows:

$$S = \left(P1\,\tilde{w}_{11} + P2\,\tilde{w}_{22} + P3\,\tilde{w}_{33}\right)$$
$$= .5\left(P1\left(w_{11}^{(1)} + w_{11}^{(2)}\right) + P2\left(w_{22}^{(1)} + w_{11}^{(2)}\right) + P3\left(w_{22}^{(1)} + w_{22}^{(2)}\right)\right) \quad (16)$$

From the "equitability" of S1 one has,

$$S1 = P1\,w_{11}^{(1)} + (P2 + P3)\,w_{22}^{(1)} = 1 \quad (17)$$

and from the "equitability" of S2 one has,

$$S2 = (P1 + P2)\,w_{11}^{(2)} + P3\,w_{22}^{(2)} = 1 \quad (18)$$

These relations when introduced in the expression (16) yield unity for S, as was to be demonstrated.

It may be useful to give some examples of the three-class equitable score weighting matrix as determined by the probability of occurrence of each of the three classes - P1, P2, and P3.

#1. P1 = P2 = P3 = 1/3 :

1.25	-.25	-1.00
-.25	.50	- .25
-1.00	-.25	1.25

Notice that #1. gives the scoring matrix in Gandin and Murphy (their eq. 26.)

#2. P1=.5 P2=P3=.25 :

.67	-.33	-1.00
-.33	.67	0.00
-1.00	0.00	2.00

It is interesting to note that in #2. no weight (positive or negative) is given for two of the nine possible combinations of forecast and observation.

#3. P1=.1 P2=.3 P3=.6 :

5.25	.25	-1.00
.25	.81	-.44
-1.00	-.44	.39

#4. P1=.1 P2=.2 P3=.7 :

5.67	.67	-1.00
.67	1.22	-.44
-1.00	-.44	.39

Cases #3. and #4. are examples involving unbalanced choices of the classes in terms of their probability of occurrence. Note that it is possible for a "near miss" of the "rare" event to receive a more positive weight than a "hit" of a relatively much higher probability event. That this is not at odds with the basic concept of an equitable score follows from the discussion in section 4. of Gandin and Murphy.

#5. P1=.01 P2=.1 P3=.89 :

53.55	3.55	-1.00
3.55	4.05	-.49
-1.00	-.49	.07

Example #5. corresponds to a case in which one class of the event is extremely unlikely. It seems somewhat unreasonable to produce such a selection of classifications; nonetheless, the equitable score weights seem reasonable, and are designed so that the forecaster is appropriately rewarded for attempts to forecast the rare event. The reward is not restricted solely to a precise hit, but is also provided for an imperfect, but suggestive, forecast.

3 FOUR CLASS TABLE

The four class table is assumed to involve a mutually exclusive and exhaustive partition of the predicted event into four classes; it is also assumed that the classes - C1, C2, C3, C4 - bear an ordered arrangement. For application to quantitative precipitation forecasting these conditions are natural, for example:

C1	precip < .01 "	C2	.01" <= precip < .25"
C3	.25" <= precip < .50"	C4	precip >= .50"

A four-class contingency table will have the form:

		Forecast				
		C1	C2	C3	C4	sum
Observed	C1	n11	n12	n13	n14	p1
	C2	n21	n22	n23	n24	p2
	C3	n31	n32	n33	n34	p3
	C4	n41	n42	n43	n44	p4
	sum	q1	q2	q3	q4	N

Derivation of an equitable skill score weight matrix proceeds as in section 3. First, the four-class table is partitioned into three, three-class tables:

Table 1

		Forecast			
		C1	C2	(C3 or C4)	sum
Observed	C1	n11	n12	(n13+n14)	p1
	C2	n21	n22	(n23+n24)	p2
	(C3 or C4)	(n31+n41)	(n32+n42)	(n33+n43+n34+n44)	(p3+p4)
	sum	q1	q2	(q3+q4)	N

Table 2

		Forecast			
		C1	(C2 or C3)	C4	sum
Observed	C1	n11	(n12+n13)	n14	p1
	(C2 or C3)	(n21+n31)	(n22+n32+n23+n33)	(n24+n34)	(p2+p3)
	C4	n41	(n42+n43)	n44	p4
	sum	q1	(q2+q3))	q4	N

Table 3

		Forecast			
		(C1 or C2)	C3	C4	sum
	(C1 or C2)	(n11+n12+n21+n22)	(n13+n23)	(n14+n24)	(p1+p2)
Observed	C3	(n31+n32)	n33	n34	p3
	C4	(n41+n42)	n43	n44	p4
	sum	(q1+q2)	q3	q4	N

Define the probabilities:

$$P1 = E\left\{\frac{p1}{N}\right\}, P2 = E\left\{\frac{p2}{N}\right\}, P3 = E\left\{\frac{p3}{N}\right\}, P4 = E\left\{\frac{p4}{N}\right\}$$

$$Q1 = E\left\{\frac{q1}{N}\right\}, Q2 = E\left\{\frac{q2}{N}\right\}, Q3 = E\left\{\frac{q3}{N}\right\}, Q4 = E\left\{\frac{q4}{N}\right\}$$

where E{ } means the expected value.

Using the results in section 2, <u>equitable</u> skill scores, S1, S2 and S3, may be written for each of the three, three class tables. By setting the <u>equitable</u> score for the four-class table equal to the arithmetic average of S1, S2 and S3, the elements of the four-class weight matrix,

$$\begin{pmatrix} \hat{w}_{11} & \hat{w}_{12} & \hat{w}_{13} & \hat{w}_{14} \\ \hat{w}_{12} & \hat{w}_{22} & \hat{w}_{23} & \hat{w}_{24} \\ \hat{w}_{13} & \hat{w}_{23} & \hat{w}_{33} & \hat{w}_{34} \\ \hat{w}_{14} & \hat{w}_{24} & \hat{w}_{34} & \hat{w}_{44} \end{pmatrix}$$

may be determined, in terms of the set of weights of the three-class tables,

$$\hat{w}_{11} = \frac{\left(w_{11}^{(1)} + w_{11}^{(2)} + w_{11}^{(3)}\right)}{3} \qquad \hat{w}_{22} = \frac{\left(w_{11}^{(1)} + w_{22}^{(2)} + w_{22}^{(3)}\right)}{3}$$

$$\hat{w}_{33} = \frac{\left(w_{22}^{(1)} + w_{22}^{(2)} + w_{33}^{(3)}\right)}{3} \qquad \hat{w}_{44} = \frac{\left(w_{33}^{(1)} + w_{33}^{(2)} + w_{33}^{(3)}\right)}{3}$$

$$\hat{w}_{12} = \frac{\left(w_{11}^{(1)} + w_{12}^{(2)} + w_{12}^{(3)}\right)}{3} \qquad \hat{w}_{13} = \frac{\left(w_{12}^{(1)} + w_{12}^{(2)} + w_{13}^{(3)}\right)}{3}$$

$$\hat{w}_{14} = \frac{\left(w_{13}^{(1)} + w_{13}^{(2)} + w_{13}^{(3)}\right)}{3} \qquad \hat{w}_{23} = \frac{\left(w_{12}^{(1)} + w_{22}^{(2)} + w_{23}^{(3)}\right)}{3}$$

$$\hat{w}_{24} = \frac{\left(w_{13}^{(1)} + w_{23}^{(2)} + w_{23}^{(3)}\right)}{3} \qquad \hat{w}_{34} = \frac{\left(w_{23}^{(1)} + w_{23}^{(2)} + w_{33}^{(3)}\right)}{3}$$

or, in terms of the probabilities, P1,P2,P3 and P4 :

$$\hat{w}_{11} = \frac{1}{3}\left(\frac{1.-P1}{P1} + \frac{1.-(P1+P2)}{(P1+P2)} + \frac{P4}{1.-P4}\right)$$

$$\hat{w}_{22} = \frac{1}{3}\left(\frac{P1}{1.-P1} + \frac{1.-(P1+P2)}{(P1+P2)} + \frac{P4}{1.-P4}\right)$$

$$\hat{w}_{33} = \frac{1}{3}\left(\frac{P1}{1.-P1} + \frac{1.-(P3+P4)}{(P3+P4)} + \frac{P4}{1.-P4}\right)$$

$$\hat{w}_{44} = \frac{1}{3}\left(\frac{P1}{1.-P1} + \frac{1.-(P3+P4)}{(P3+P4)} + \frac{1.-P4}{P4}\right)$$

$$\hat{w}_{12} = \hat{w}_{21} = \frac{1}{3}\left(-1. + \frac{1.-(P1+P2)}{P1+P2} + \frac{P4}{1.-P4}\right)$$

$$\hat{w}_{13} = \hat{w}_{31} = \frac{1}{3}\left(-2. + \frac{P4}{1.-P4}\right)$$

$$\hat{w}_{14} = \hat{w}_{41} = -1.$$

$$\hat{w}_{23} = \hat{w}_{32} = \frac{1}{3}\left(\frac{P1}{1.-P1} + \frac{P4}{1.-P4} - 1.\right)$$

$$\hat{w}_{24} = \hat{w}_{42} = \frac{1}{3}\left(\frac{P1}{1.-P1} - 2.\right)$$

$$\hat{w}_{34} = \hat{w}_{43} = \frac{1}{3}\left(\frac{P1}{1.-P1} + \frac{P1+P2}{(P3+P4)} - 1.\right)$$

The proof of the equitability of this four class weight matrix follows the same method used for the three class weight matrix in section 2, and doesn't warrant the use of additional space here. (A BASIC language program is given in an appendix which the reader may use to satisfy any misgivings on this point.)

Some examples are given below for equitable weights for four class contingency tables with the indicated probabilities of occurrence for each class: P1,P2,P3 and P4.

#1. P1=P2=P3=P4=.25:

1.44	0.11	-0.56	-1.00
0.11	0.56	-0.11	-0.56
-0.56	-0.11	0.56	0.11
-1.00	-0.56	0.11	1.44

#2. P1=P4=.10, P2=P3=.40:

3.37	0.04	-0.63	-1.00
0.04	0.41	-0.26	-0.63
-0.63	-0.26	0.41	0.04
-1.00	-0.63	0.04	3.37

#3. P1=.10, P2=.20, P3=.30, P4=.40:

4.00	0.67	-0.44	-1.00

0.67	1.04	-0.07	-0.63
-0.44	-0.07	0.40	-0.15
-1.00	-0.63	-0.15	0.68

#4. P1=.50, P2=P3=.24, P4=.02

0.46	-0.21	-0.66	-1.00
-0.21	0.46	0.01	-0.33
-0.66	0.01	1.29	0.95
-1.00	-0.33	0.95	17.62

#5 P1=.50, P2=P3=.249, P4=.002

0.45	-0.22	-0.67	-1.00
-0.22	0.45	0.00	-0.33
-0.67	0.00	1.33	0.99
-1.00	-0.33	0.99	167.66

Examples #4 and #5 are given because they illustrate how the four class scoring matrix "approaches" that of its three class counterpart when one of the four classes has a probability of occurring that approaches zero.

4 APPLICATION TO ETA FORECASTS

The sample of ETA model forecasts of 24 hour accumulation of precipitation was produced during the period April 28 to May 31 1991. Only the first period forecasts, hours 0 to 24, and only forecasts initiated at 1200GMT were used. The verification data were the observations made by both cooperative and regular components of the NWS River Forecast Network. The observations cover the day, beginning and ending at 1200 GMT. Unfortunately, only observations of occurrence of precipitation are reported by the cooperative component of the network; non-occurrence is not reported.

The observed data are applied to the forecast model's gridpoints, by assigning to the gridpoint the arithmetic mean of the observations lying within the cell surrounding the gridpoint. The ETA model uses a grid which has separation interval of approximately 80 km. Only gridpoints where there is a possibility of an observation being reported are included in the verification; there are 1060 such points over the contiguous United States.

Four classes were used in the verification,

C1 : R < .01"   C2 : .01 <= R < .50"

C3 : .50"<= R < 1.00"   C4 : 1.00" <= R

where R stands for the amount of precipitation, forecast or observed. The collection of data was missed on three days; so the sample involves just 29 days.

Note that the four classes are partitioned by three threshholds - .01", .50" and 1.00". It can be shown (section 6) that the <u>equitable</u> score for the four class contingency table may be constructed from the arithmetic average of the <u>equitable</u> scores for each of three two class tables, one for each of the threshholds. The data for each day and each threshhold are given in appendix II.

The contingency tables constructed from the average of the daily entries in each cell of the table are:

		FORECAST		
		YES	NO	SUM
	YES	239.5	155	394.5
OBSERVED	NO	142.5	523	665.5
	SUM	382	678	1060.

Contingency table for .01" threshhold

		FORECAST		
		YES	NO	SUM
	YES	35.	52.4	87.4
OBSERVED	NO	45.8	926.8	972.6
	SUM	80.8	979.2	1060.

Contingency table for .50" threshhold

		FORECAST		
		YES	NO	SUM
	YES	7.8	18.9	26.7
OBSERVED	NO	18.0	1015.3	1033.3
	SUM	25.8	1034.2	1060.

Contingency table for 1.00" threshhold

For the .01" threshhold the probability of occurrence is .372, so the equitable score weights are 1.69 and .59 for the successful forecasts of occurrence and non-occurrence, respectively. The resulting equitable score is .39.

For the 0.5" threshhold the probability of occurrence is .082, so the equitable score weights are 11.1 and .09 for the successful forecasts of occurrence and non-occurrence, respectively. The resulting equitable score is .35.

For the 1.0" threshhold the probability of occurrence is .025, so the equitable score weights are 38.7 and .025 for the successful forecasts of occurrence and non-occurrence, respectively. The resulting equitable score is .27.

Thus the equitable score for the full four class table is 0.34.

Other statistics for this data set are:

Threshhold	TS	Bias	P(F|O)	P(O|F)	GTS	TSS
.01"	.44	.97	.61	.63	.14	.29
.50"	.26	.92	.40	.43	.20	.23
1.00"	.17	.96	.29	.30	.16	.16
Average	.29	.95	.43	.45	.17	.23

where in addition to the previously defined threat score (TS) and bias: P(F|O) is the conditional probability that the event was forecast given that it was observed; P(O|F) is the conditional probability that the event was observed given that it was forecast; GTS is the Gilbert threat score (Schaefer, 1990) defined by,

$$GTS = \frac{(H1 - E\{H1\})}{F + O - (H1 - E\{H1\})}$$

in which E{H1} is the number of successful forecasts of the occurrence of the event to be expected due to chance. The calculation of E{H1} is done using the formula,

$$E\{H1\} = \frac{F\ O}{N} \ ;$$

and TSS is the skill score defined in terms of the threat score by the relation,

$$TSS = \frac{TS - TS_r}{1. - TS_r}$$

where

$$TS_r = \frac{E\{H1\}}{F + O - E\{H1\}}$$

is the threat score for a random forecast.

While it is true that inspection of the full four class contingency table is necessary to comprehend the full character of the joint distribution of forecasts and observations, it is desirable to produce a statistic that digests the full set of information and provides an "equitable" assessment of performance. From the statistical data, presented above, the GTS, Skill and Equitable scores seem to possess the desirable property of small, inter-threshhold variability, when compared to the TS and the two conditional probabilities. This is so because they take into account the difference in difficulty associated with forecasts at the different threshholds. Scores with this property lend themselves to graphical presentation without the tendency for over-emphasis on the lower threshholds, that is evident in plots of the Threat Score versus threshhold.

5 APPLICATION TO NGM and ETA FORECASTS

To gain insight into the applicability of the equitable score to comparison of model forecasts, the equitable score and other statistics were computed for a small sample of quantitative precipitation forecasts made, during June 1991, by two different regional models, the NGM (Phillips, 1976) and the ETA (Messinger, Black and Janjic, 1988.) Because the models use different computational grids, there is a difference in the number of points used in the verification, but the verification area, the contiguous United States, is essentially the same. The source of verification data is the same as outlined in the previous section.

The contingency tables are based on the four classes used in section 4, (C1: R<.01", C2: .01"<=R<.5". C3: .5"<=R<1." and C4: R>1.") The sample consists of sixteen daily forecasts made between 7 June and 23 June (the 22nd is missing.)

Since the sample is so small, it is evident that no conclusions may be drawn from this data regarding the merits of one model versus the other. The two models do have somewhat different characteristics, and it is of interest to learn if the equitable score provides a way to account rationally for the these differences.

In presenting the average contingency tables, the entries are given as decimal fractions of the total number of verification points. This avoids the confusion which might arise from the different number of verification points (1261 for the NGM; 1060 for the ETA model.)

		Forecast		
		Yes	No	Sum
Observed	Yes	.24	.06	.30
	No	.23	.47	.70
	Sum	.47	.53	1.00

Contingency table for .01" NGM forecasts.

		Forecast		
		Yes	No	Sum

Observed	Yes	.18	.14	.32
	No	.12	.56	.68
	Sum	.30	.70	1.00

Contingency table for .01" ETA forecasts.

		Forecast		
		Yes	No	Sum
Observed	Yes	.02	.05	.07
	No	.06	.87	.93
	Sum	.08	.92	1.00

Contingency table for .5" NGM forecasts.

		Forecast		
		Yes	No	Sum
Observed	Yes	.02	.05	.07
	No	.04	.89	.93
	Sum	.06	.94	1.00

Contingency table for .5" ETA forecasts.

		Forecast		
		Yes	No	Sum
Observed	Yes	.002	.018	.02
	No	.008	.972	.98
	Sum	.010	.990	1.00

Contingency table for 1." NGM forecasts.

		Forecast		
		Yes	No	Sum
Observed	Yes	.003	.017	.02
	No	.020	.960	.98
	Sum	.023	.977	1.00

Contingency table for 1." ETA forecasts.

The probability of occurrence of each of the four classes differs slightly between the ETA and NGM data sets. For the ETA model data set, the probabilities of occurrence in the four classes are .68, .25, .05 and .02. For the NGM data set, the probabilities are: .70, .23, .05 and .02.

The equitable score was calculated based on the three two-class tables associated with the threshholds separating the four classes. The resulting four-class equitable score for the NGM is 0.29 and for the ETA model is .25.

It was surprising that the ETA model's score for this period in June was so much less than the .34 score found in the May data set, analyzed in section 4. Examining the contributions to the four-class score made by the three threshholds shows that the degradation in score is caused by performance at the larger amounts of precipitation. This is shown in the following table

	Threshhold .01"	.5"	1."
May	.39	.35	.27
June	.40	.22	.11

Table Contributions to the ETA model's Equitable score from the three threshholds.

The probability of occurrence of precipitation exceeding each of these threshholds during the two data collections is tabulated below.

	Threshhold .01"	.5"	1."
May	.37	.082	.025
June	.31	.071	.019

Table Probability of occurrence of precipitation exceeding the three threshholds.

The forecast model's frequency of forecasting precipitation amounts exceeding the several threshholds may also be expressed in terms of probability measures, as in the following table.

	Threshhold .01"	.5"	1."
May	.36	.075	.024
June	.30	.063	.022

Table Probability of prediction of precipitation exceeding the three threshholds.

It does not seem likely that the small variations in observed and predicted probabilities of occurrence would distort the computations of the equitable score. This leaves the conclusion that the accuracy of the ETA model's forecasts, during the June data period, were not affected by unusual, systematic biases but by other factors, that the precipitation verification statistics cannot elucidate.

Attention may be turned to a comparison of the equitable score with three other statistical estimates of performance. These are the previously defined threat score (TS), bias (B), conditional probabilities (P(F|O) and P(O|F)), Gilbert threat score (GTS), threat skill score (TSS), and the Gandin-Murphy equitable skill score (GMS.) Also tabulated is the threat score for a random forecast (TSR.) Although most of these scores can be calculated day-by-day, only values for the average contingency table will be considered.

Threshhold	TS	B	P(F|O)	P(O|F)	GTS	TSS	GMS	TSR
.01"	.42	.94	.57	.61	.16	.29	.40	.18
.50"	.16	.87	.26	.30	.12	.14	.22	.04
1.00"	.06	1.15	.13	.11	.05	.05	.11	.01

Table Statistics for ETA model precipitation forecasts at three threshholds for June data set.

Threshhold	TS	B	P(F|O)	P(O|F)	GTS	TSS	GMS	TSR
.01"	.44	1.57	.79	.50	.14	.28	.55	.22
.50"	.18	1.09	.31	.29	.13	.14	.26	.04
1.00"	.06	.47	.08	.16	.05	.05	.07	.01

Table Statistics for NGM model precipitation forecasts at three threshholds for June data set.

The characteristic difference between the two models' forecasts is most clearly manifested the Bias score and the associated threat score for a random forecast (TSR.) The NGM has a large bias for small amounts and a very small bias for large amounts of precipitation. The ETA model manifests the opposite trend but with a smaller variation.

The equitable score for the four class table judged the NGM performance to be superior the ETA. Yet, when the large contribution due to the heavily biased low threshhold is noted, it seems that this judgment is questionable. Indeed, the Gilbert threat, score when averaged over the three threshholds, gives the nod to the ETA models. The threat skill score suggests that when the variation in bias is taken into account both models have almost identical skill compared to random forecasts made using the associated model's estimate of the probabiltiy of occurrence in each class.

This data set is so small that only extremely tentative conclusions are warranted, but it seems reasonable to conclude that the equitable score will not be a panacea for the problem of comparing the performance of models possessing significantly different bias characteristics. It also seems appropriate to include the threat skill score in future model comparisons.

6 ADDITIONAL REMARKS ON THE EQUITABLE SCORE

It is one of the attractions of the threat score that it can be given a simple "geometric" interpretation in terms of the intersection and union of the forecast and observed data. For the interpretation of the "meaning" of the equitable score, the following analysis is helpful.

For the two class contingency table,

		Forecast		
		Yes	No	Sum
	Yes	H1	M1	O
Observed	No	M2	H2	N-O
	Sum	F	N-F	N

the equitable score S is written in terms of the entries H1, H2, M1 and M2, together with two weights that are defined in terms of the climatological expectency of O/N, which may be denoted by P. The expression is,

$$S = \left(\frac{1.-P}{P}\right)\frac{H1}{N} + \left(\frac{P}{1.-P}\right)\frac{H2}{N} - \frac{M1}{N} - \frac{M2}{N}.$$

If the following notational simplifications are employed,

$$h_1 = \frac{H1}{O} \qquad h_2 = \frac{H2}{N-O}$$

$$\mu = \frac{O}{N} \qquad \nu = \frac{N-O}{N} = 1.-\mu$$

and M1 and M2 are replaced by O-H1 and N-O-H2, S may be written as

$$S = \left(1.+\frac{1-P}{P}\right)\mu h_1 + \left(1.+\frac{P}{1.-P}\right)\nu h_2 - 1$$

It is evident that P is the expected value of μ which can be written $P = E\{\mu\}$, with the corrolary, $(1-P) = E\{\nu\}$. So the expression for the equitable score can be put into the form,

$$S = \frac{\mu}{E\{\mu\}}h_1 + \frac{\nu}{E\{\nu\}}h_2 - 1.$$

When the data set has the property that μ and ν are close to their expected values, then the equitable score is approximately given by

$$S = (h_1 + h_2 - 1.)$$

So provided the individual case does not depart too much from the climatological expectency of the event's occurrence, the <u>equitable</u> score may be interpreted as the fraction of correct forecasts of the event's occurrence plus the fraction of correct forecasts of the event's non-occurrence minus unity! To optimize this score it is necessary to accurately predict both occurrence <u>and</u> non-occurrence of the event.

In section 4. it was noted that the evaluation of the <u>equitable</u> score for the four-class table reduced to the evaluation of three, two-class table scores -- one for each of the three threshholds present in the four class table. The explanation for this can be seen in the following schematic :

$$S$$

$$C1 \qquad C2 \qquad C3 \qquad C4$$

$$S_1 \qquad S_2 \qquad S_3$$

$$1.\ C1 \oplus C2\ \ C3\ \ C4 \qquad 2.\ C1\ \ C2 \oplus C3\ \ C4 \qquad 3.\ C1\ \ C2\ \ C3 \oplus C4$$

$$S_{1,1} \qquad S_{2,1} \qquad S_{3,1}$$

$$1a.\ C1 \oplus C2 \oplus C3\ \ C4 \quad 2a.\ C1 \oplus C2 \oplus C3\ \ C4 \quad 3a.\ C1 \oplus C2\ \ C3 \oplus C4$$

$$S_{1,2} \qquad S_{2,2} \qquad S_{3,2}$$

$$1b.\ C1 \oplus C2\ \ C3 \oplus C4 \quad 2b.\ C1\ \ C3 \oplus C3 \oplus C4 \quad 3b.\ C1\ \ C2 \oplus C3 \oplus C4$$

Briefly, the score S for the four-class table reduces to the average of the three scores, S_1,S_2, &S_3 for the three itemized partitions of the four-class table into three-class tables. In turn 1a., 1b. 2a., 2b., 3a., and 3b. are partitions of the three-class tables into two-class tables with the associated scores, $S_{1,1}$, $S_{1,2}$, $S_{2,1}$, $S_{2,2}$, $S_{3,1}$, and $S_{3,2}$.

But the set of six, two-class tables contains only three unique partitions, each of which is associated with one of the three threshholds of the original four-class table. Thus, in evaluating the <u>equitable</u> score S, it suffices to evaluate $S_{1,1}$, $S_{1,2}$, and $S_{3,2}$, as was done in section 4.

7 SUMMARY AND CONCLUSIONS

The equitable skill score proposed by Gandin and Murphy (1991) for evaluating the performance of categorical forecasts was applied to small samples of numerical model predictions of quantitative precipitation. The application of the score to multiple class contingency tables was found to reduce to the computation of scores for a set of two-class tables, one for each threshhold in the multiclass table. This was demonstrated for three and four class tables, and is conjectured, but not proven, to be generally true.

It was expected that the equitable score's theoretical property of insensitivity to biased, random forecasting strategies might prove useful in comparing forecast models that possess different characteristics in the bias of their quantitative precipitation forecasts. In a reported test, this expectation was not completely fulfilled, and the alternative threat skill score statistic seemed to be somewhat better behaved.

It was hoped, perhaps naively, that the application of the equitable score to a multiclass table of quatitative precipitation forecasts and observations would lessen the need to examine scores for several threshholds of precipitation. Experience, gained in tests reported above, suggests that this hope will not be fulfilled. In part, because the computation of the score for the multiclass table can be carried-out considering the set of two-class partitions of the event associated with each of the several threshholds embedded in the multiclass table, it is unlikely that the contributions made to the final score by each partition can be overlooked.

It is expected that future efforts in the verification of precipitation forecasts will use the equitable score. It is of theoretical interest to prove that the computation of an equitable score for a general K-class table can be reduced to computing K-1 two-class table scores, and the corrolary that the weight matrix for the K-class table may be built-up from the weights of the K-1 two class tables.

8 REFERENCES

Gandin, L.S. and A.H. Murphy, 1991: Equitable Skill Scores for Categorical Forecasts. To appear in Monthly Weather Review, American Meteorological Society, Boston, MA.

Mesinger, F., T.L. Black and Z.I. Janjic, 1988: A Summary of the NMC Step-Mountain (ETA) Coordinate Model. pp 91-98, Proceedings Workshop on Limited Area Modeling Intercomparison, 15-18 Nov. 1988, Boulder CO., National Center of Atmospheric Research, Boulder, CO.

Phillips, N.A. 1976: The Nested Grid Model. NOAA Technical Report NWS22 , 80pp., National Weather Service, Washington, D.C.

Schaefer, J.T., 1990: The Critical Success Index as an Indicator of Warning Skill. Weather and Forecasting, 5, 570-575, American Meteorological Society, Boston, MA.

9 APPENDIX I BASIC PROGRAM

```
10  REM BASIC LANGUAGE PROGRAM TO TEST EQUITABILITY
20  REM OF 4-CLASS WEIGHT MATRIX
30  PRINT "INPUT P1"
40  INPUT P1
50  PRINT "INPUT P2"
60  INPUT P2
70  PRINT "INPUT P3"
80  INPUT P3
90  P4=1 - (P1+P2+P3)
100 PRINT "PROBABILITIES OF EVENT CLASSES
110 PRINT USING "P1=#.##  P2=#.##  P3=#.##  P4=#.##";P1,P2,P3,P4
120 W11=(1/3)*( (1-P1)/P1 + (P3+P4)/(P1+P2) + P4/(1-P4) )
130 W22=(1/3)*( P1/(1-P1) + (P3+P4)/(P1+P2) + P4/(1-P4) )
140 W33=(1/3)*( P1/(1-P1) + (P1+P2)/(P3+P4) + P4/(1-P4) )
150 W44=(1/3)*( P1/(1-P1) + (1-P3-P4)/(P3+P4) + (1-P4)/P4 )
160 W12=(1/3)*( -1. + (P3+P4)/(P1+P2) + P4/(1-P4) ) : W21=W12
170 W13=(1/3)*( -2. + P4/(1-P4) : W31=W13
180 W14=(1/3)*( -3)  :  W41=W14
190 W23=(1/3)*( P1/(1-P1) + P4/(1-P4) -1) : W32=W23
200 W24=(1/3)*( P1/(1-P1) -2 )  : W42=W24
210 W34=(1/3)*( P1/(1-P1) + (P1+P2)/(P3+P4) -1)  : W43=W34
220 PRINT "WEIGHT MATRIX"
230 PRINT USING "##.##   ##.##   ##.##   ##.## ";W11,W12,W13,W14
240 PRINT USING "##.##   ##.##   ##.##   ##.## ";W21,W22,W23,W24
250 PRINT USING "##.##   ##.##   ##.##   ##.## ";W31,W32,W33,W34
260 PRINT USING "##.##   ##.##   ##.##   ##.## ";W41,W42,W43,W44
270 REM  NOW TEST FOR RANDOM CONTINGENCY TABLE
280 PRINT "INPUT Q1"
290 INPUT Q1
300 PRINT "INPUT Q2"
310 INPUT Q2
320 PRINT "INPUT Q3"
330 INPUT Q3
340 Q4= 1. - (Q1+Q2+Q3)
350 PRINT USING "FCST PROBS Q1=.## Q2=.## Q3=.##
Q4=.##";Q1,Q2,Q3,Q4
360 REM GENERATE ELEMENTS OF 4 CLASS RANDOM CONTINGENCY TABLE
370 E11=P1*Q1 : E12=P1*Q2 : E13=P1*Q3 : E14=P1*Q4
380 E21=P2*Q1 : E22=P2*Q2 : E23=P2*Q3 : E24=P2*Q4
390 E31=P3*Q1 : E32=P3*Q2 : E33=P3*Q3 : E34=P3*Q4
400 E41=P4*Q1 : E42=P4*Q2 : E43=P4*Q3 : E44=P4*Q4
410 REM CALCULATE EQUITABLE SCORES
420 REM SK FOR ONLY FORECAST CLASS K
430 REM SR FOR RANDOM TABLE
440 S1=W11*E11+W21*E21+W31*E31+W41*E41
450 S2=W12*E12+W22*E22+W32*E32+W42*E42
```

```
460 S3=W13*E13+W23*E23+W33*E33+W43*E43
470 S4=W14*E14+W24*E24+W34*E34+W44*E44
480 S= S1+ S2+ S3+ S4
490 PRINT USING "S1=#.## S2=#.## S3=#.## S4=#.##
S=#.##";S1,S2,S3,S4,S
500 REM TEST FOR PERFECT FORECAST
510 S=W11*P1 + W22*P2 + W33*P3 + W44*P4
520 PRINT USING "PERFECT FORECAST S=##.##";S
530 END
```

10 APPENDIX II DATA FOR SECTION 4

	.01"			.50"			1.0"		
	HITS	O	F	HITS	O	F	HITS	O	F
APR28	292	403	502	51	88	127	3	39	15
APR29	382	494	553	78	145	156	25	68	47
APR30	309	453	433	74	109	117	38	50	68
MAY1	170	345	293	3	33	30	0	11	8
MAY2	147	282	275	2	20	7	0	2	0
MAY3	341	461	514	65	113	117	19	46	31
MAY4	313	454	432	60	140	117	7	31	54
MAY5	336	442	418	82	136	133	28	52	54
MAY6	169	312	260	40	60	47	1	18	9
MAY7	173	339	289	38	90	71	7	38	8
MAY9	215	389	339	42	88	81	15	32	37
MAY10	183	329	361	18	37	72	2	10	20
MAY11	187	312	400	21	52	49	1	7	16
MAY12	160	285	355	8	60	38	0	4	4
MAY13	223	388	399	26	77	65	1	18	12
MAY14	159	343	298	6	69	17	0	15	6
MAY15	191	408	312	12	89	40	3	22	8
MAY16	229	403	367	33	98	95	5	31	22
MAY17	327	483	475	40	138	82	1	28	16
MAY18	321	454	507	57	131	132	6	30	41
MAY20	242	380	330	26	73	107	2	14	40
MAY21	237	372	346	13	47	61	2	11	11
MAY22	211	399	317	10	53	63	1	17	17
MAY23	241	420	367	27	70	99	10	23	48
MAY26	253	425	402	66	129	124	26	53	69
MAY27	262	423	407	61	100	130	15	27	63
MAY29	207	365	311	26	81	69	3	24	9
MAY30	217	441	366	15	106	41	3	31	8
MAY31	248	436	450	14	103	57	1	22	6